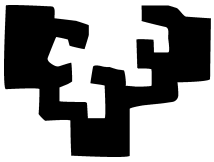

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 2

EJERCICIOS

1. Encuentra los estimadores de momentos del vector de medias y matriz de covarianzas basados en una muestra de tamaño N procedente de *cualquier* distribución multivariante.

2. Encuentra los estimadores máximo verosímiles homólogos cuando la distribución generadora de las observaciones es normal multivariante. ¿Son insesgados ambos estimadores máximo verosímiles? (Ayuda: Necesitarás derivar una función de verosimilitud respecto de toda una matriz. Puedes recurrir a resultados —que no hemos visto en clase— disponibles en libros de Análisis Multivariante, por ej. [7] o [1].)

3. Demuestra que el estadístico T^2 de Hotelling para el contraste de hipótesis sobre el vector de medias de una población es invariante frente a transformaciones lineales no singulares (es decir, demuestra que si en lugar de emplear las observaciones originales $\vec{X}_i$ $(i = 1, \ldots, N)$ para hacer el contraste empleases $\vec{Y}_i = A\vec{X}_i$ siendo A una matriz no singular, el resultado sería exactamente el mismo).

4. Para demostrar que NS^2 sigue una distribución $\mathrm{Wishart}(N-1, \Sigma)$ empleamos el siguiente procedimiento:

   a) Demostrar que si
   $$X = \begin{pmatrix} \vec{X}_1 \\ \vdots \\ \vec{X}_N \end{pmatrix} \qquad Y = UX = \begin{pmatrix} \vec{Y}_1 \\ \vdots \\ \vec{Y}_N \end{pmatrix}$$
   y U es una matriz ortogonal, entonces $Y'Y = \sum_{i=1}^N \vec{Y}_i\vec{Y}_i{}' = \sum_{i=1}^N \vec{X}_i\vec{X}_i{}' = X'X$. Además, si las filas de X se distribuyen como $N_d(\vec{0}, \Sigma)$, las filas $\vec{Y}_i{}'$ de Y se distribuyen también del mismo modo.

   b) Emplear una matriz ortogonal U especial en la transformación anterior, cuya última fila era $\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}$. Con tal matriz se tiene que $\vec{Y}_N = \sqrt{N}\,\vec{\overline{X}}$.

   c) Mostrar entonces que
   $$NS^2 = \sum_{i=1}^N \vec{X}_i\vec{X}_i{}' - N\vec{\overline{X}}\,\vec{\overline{X}}{}' = \sum_{i=1}^N \vec{Y}_i\vec{Y}_i{}' - \vec{Y}_N\vec{Y}_N{}'$$

   d) De la igualdad anterior, concluir que NS^2 es una suma de "cuadrados" $\vec{Y}_i\vec{Y}_i{}'$ de vectores normales con matriz de covarianzas Σ, y por consiguiente sigue una distribución $\mathrm{Wishart}(N-1, \Sigma)$.

En ningún momento, sin embargo, se ha demostrado que cuando los $\vec{X}_i$ tienen vector de medias arbitrario los $\vec{Y}_i$ $(i = 1, \ldots, N-1)$ tengan vector de medias nulo, como es necesario para que lo afirmado en (4*d*) sea correcto. Completa la demostración (Ayuda: fíjate en la estructura de la matriz U; si la última fila es la que se ha dicho, ¿qué ocurre con las restantes? ¿Qué sucede cuando una de las restantes filas multiplica a un vector todas cuyas coordenadas son iguales?).

5. En clase se mencionó que el contraste T^2 de Hotelling para la hipótesis $H_0 : \vec{\mu} = \vec{\mu}_0$ es notablemente robusto ante desviaciones de la hipótesis de normalidad multivariante. Haz una pequeña simulación para explorar hasta donde llega esta robustez. Repite cien veces, guardando los resultados cada vez, lo siguiente:

   a) Genera 200 vectores $\vec{X}_1, \ldots, \vec{X}_{200}$ de dimensión 5 procedentes de una distribución *no* normal.

   b) Obtén vectores transformados linealmente, $\vec{Y}_1, \ldots, \vec{Y}_{200}$, de modo que tengan cierta correlación y vector de medias $\vec{\mu}_0$ conocido.

   c) Estima vector de medias y matriz de covarianzas, y a continuación el estadístico T^2 para contraste de la hipótesis y su transformación en un estadístico con distribución $\mathcal{F}$ de Snedecor. Guarda este último resultado en cada iteración.

   Al finalizar, tendrás 100 valores del estadístico de contraste para la hipótesis $H_0 : \vec{\mu} = \vec{\mu}_0$ (cierta, por construcción de las observaciones). Si las observaciones hubieran sido normales multivariantes, dichos 100 valores procederían de una distribución $\mathcal{F}$ de Snedecor con grados de libertad adecuados. Al fallar el supuesto de normalidad, la distribución no será ya tal. Compara la probabilidad teórica de rechazo de H_0 bajo la hipótesis de normalidad (= α, nivel de significación) con el porcentaje empírico de rechazos en las 100 repeticiones del experimento.

6. La Tabla 1 recoge datos correspondientes a composición química de 26 restos de alfarería romana recogidas en cuatro diferentes lugares de Gales. Además de la variable de localización, `Sitio`, se midieron en cada resto los porcentajes de óxidos de varios metales, según se detalla en la siguiente clave:

Variable	Descripción
Al	Porcentaje de óxido de aluminio.
Fe	Porcentaje de óxido de hierro.
Mg	Porcentaje de óxido de magnesio.
Ca	Porcentaje de óxido de calcio.
Na	Porcentaje de óxido de sodio.
Sitio	Llanederyn (L), Island Thorns (I) Caldicot (C), Ashley Rails (A)

   El examinar las similitudes o diferencias de las composiciones químicas tiene interés porque arrojaría alguna luz sobre los flujos de comercio (y de técnicas y conocimientos) de unos lugares a otros.

   a) Estima los vectores de medias correspondientes a cada uno de los cuatro lugares.

   b) Estima las matrices de covarianzas que puedas correspondientes a los distintos lugares.

   c) Bajo el supuesto de matriz de covarianzas común, contrasta lo que es directamente de interés: igualdad de vectores de medias en las respectivas subpoblaciones. Explica las dificultades que encuentras.

   d) Sobre las dificultades que hayas mencionado en el apartado anterior, ¿qué otros inconvenientes ves a tu modo de actuar en el apartado anterior? (Ayuda: si para contrastar igualdad de medias haces contrastes por parejas al nivel de significación α, ¿qué ocurre con el nivel de significación conjunto?)

7. En la *dataframe* `Sitka89` (lee con `read.table`) tienes datos correspondientes a 79 árboles, medidos en ocho ocasiones a lo largo de 1989. De ellos, 54 fueron cultivados en cámaras enriquecidas en ozono, y 25 fueron controles (cultivados en el medio ambiente, sin tratamiento especial). Los significados de las variables son:

   Haz un contraste de igualdad de vectores de medias entre las dos poblaciones (árboles tratados y controles).

Cuadro 1: Composición química de restos de alfarería romana hallados en Gales. Los datos son porcentajes de óxidos de distintos metales.

Al	Fe	Mg	Ca	Na	Sitio	Al	Fe	Mg	Ca	Na	Sitio
14.4	7.00	4.30	0.15	0.51	L	12.5	6.44	3.94	0.22	0.23	L
13.8	7.08	3.43	0.12	0.17	L	11.8	5.44	3.94	0.30	0.04	C
14.6	7.09	3.88	0.13	0.20	L	11.6	5.39	3.77	0.29	0.06	C
11.5	6.37	5.64	0.16	0.14	L	18.3	1.28	0.67	0.03	0.03	I
13.8	7.06	5.34	0.20	0.20	L	15.8	2.39	0.63	0.01	0.04	I
10.9	6.26	3.47	0.17	0.22	L	18.0	1.50	0.67	0.01	0.06	I
10.1	4.26	4.26	0.20	0.18	L	18.0	1.88	0.68	0.01	0.04	I
11.6	5.78	5.91	0.18	0.16	L	20.8	1.51	0.72	0.07	0.10	I
11.1	5.49	4.52	0.29	0.30	L	17.7	1.12	0.56	0.06	0.06	A
13.4	6.92	7.23	0.28	0.20	L	18.3	1.14	0.67	0.06	0.05	A
12.4	6.13	5.69	0.22	0.54	L	16.7	0.92	0.53	0.01	0.05	A
13.1	6.64	5.51	0.31	0.24	L	14.8	2.74	0.67	0.03	0.05	A
12.7	6.69	4.45	0.20	0.22	L	19.1	1.64	0.60	0.10	0.03	A

Fuente: Datos de la colección para la docencia en `StatLib`, originalmente procedentes de [6]. Puedes encontrar los datos en el fichero `pottery.dat`, en la ubicación habitual.

Cuadro 2: Variables de la dataframe `Sitka89`

Variable	**Descripción**
`size`	Producto altura por diámetro al cuadrado, en escala logarítmica.
`Time`	Momento de la medida (en días desde 1 Enero 89).
`tree`	Identificador del árbol.
`treat`	Tratamiento recibido (ozono o control)

Fuente: Datos en la biblioteca de funciones para R `MASS`, aneja a [9]. Puedes encontrar los datos en el fichero `Sitka89.dat`, en la ubicación habitual.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Al margen de tus apuntes de clase puedes mirar las secciones relevantes de cualquiera de los muchos manuales a tu disposición en Biblioteca (clasificados en 519.237, segunda planta). En particular puedes consultar [8], [3] y [7], [4] o [2] Referencias de nivel bastante más alto son [1] y [5].

2. Hay una diferencia notable entre los datos de alfarería romana y los de crecimiento de árboles. En estos últimos hay más estructura: un mismo árbol es observado en diferentes momentos de tiempo, lo que introduce restricciones (por ejemplo, los árboles no "encogen", de manera que las medidas correspondientes a un mismo árbol son monótonas). Son *datos longitudinales*.

   Esta mayor estructura sugiere la posibilidad de una parametrización más parca. En el ejemplo de los árboles, en que además se tienen los momentos de las medidas y el intervalo entre ellos por consiguiente, podríamos modelizar la matriz de covarianzas entre medidas con menos parámetros —por ejemplo, con una única tasa de crecimiento—. No es éste el objeto de la tarea. Sobre datos longitudinales encontrarás bastante bibliografía.

Referencias

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.

[2] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.

[3] J.J. Hair, R.E. Anderson, R.L. Tatham, and W.C. Black. *Multivariate Data Analysis*. Maxwell MacMillan, New York, 1992.

[4] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[5] A. Kshirsagar. *Multivariate Analysis*. Marcel Dekker, 1978.

[6] G. Nickless, A. Tubb, and A.J. Parker. The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry*, 22:153–171, 1980.

[7] A.C. Rencher. *Methods of Multivariate Analysis*. Wiley, 1995.

[8] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.

[9] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with* S-PLUS. Springer-Verlag, New York, 1994. Signatura: 681.03.068 VEN.